

1 **Assessing host-virus co-divergence for close relatives of Merkel cell polyomavirus**
2 **infecting African great apes**
3
4 Nadège F. Madinda,^{1,2,3} Bernhard Ehlers,⁴ Joel O. Wertheim,⁵ Chantal Akoua-Koffi,⁶ Richard
5 A. Bergl,⁷ Christophe Boesch,² Dieudonné Boji Mungu Akonkwa,⁸ Winnie Eckardt,^{9,10}
6 Barbara Fruth,^{11,12} Thomas R. Gillespie,^{10,13} Maryke Gray,¹⁴ Gottfried Hohmann,² Stomy
7 Karhemere,¹⁵ Deo Kujirakwinja,¹⁶ Kevin Langergraber,^{2,17} Jean-Jacques Muyembe,¹⁵ Radar
8 Nishuli,⁸ Maude Pauly,^{1,4,§} Klara J. Petrzalkova,^{18,19,20,21} Martha M. Robbins,² Angelique
9 Todd,²² Grit Schubert,¹ Tara S. Stoinski,^{9,10} Roman M. Wittig,^{2,23} Klaus Zuberbühler,^{24,25,26}
10 Martine Peeters,^{27,28} Fabian H. Leendertz,¹ Sébastien Calvignac-Spencer^{1#}
11
12 Epidemiology of highly pathogenic microorganisms, Robert Koch Institute, Berlin,
13 Germany¹; Department of primatology, Max Planck Institute for Evolutionary Anthropology,
14 Leipzig, Germany²; Institut de Recherches en Ecologie Tropicale, Libreville, Gabon³; FG12
15 “Measles, Mumps, Rubella and Viruses affecting immunocompromised patients”, Robert
16 Koch Institute, Berlin, Germany⁴; Department of Medicine, University of California, San
17 Diego, USA⁵; Centre de Recherche pour le Développement, Université Alassane Ouattara de
18 Bouake, Bouake, Côte d’Ivoire⁶; North Carolina Zoological Park, Asheboro, USA⁷; Institut
19 Congolais pour la Conservation de la Nature, Democratic Republic of Congo⁸; Diane Fossey
20 Gorilla Fund International, Atlanta, USA⁹; Department of Environmental Sciences and
21 Program in Population Biology, Ecology and Evolution, Emory University, Druid Hills,
22 USA¹⁰; Division of Neurobiology, Ludwig-Maximilians-University, Munich, Germany¹¹;
23 Centre for Research and Conservation, Royal Zoological Society of Antwerp, Antwerp,
24 Belgium¹²; Department of Environmental Health, Rollins School of Public Health, Emory
25 University, Druid Hills, USA¹³; International Gorilla Conservation Program, Kigali,
26 Rwanda¹⁴; Institut National de Recherche Biomédicale, Kinshasa, Democratic Republic of

27 Congo¹⁵; Wildlife Conservation Society, Grauer's Gorilla Project, Democratic Republic of
28 Congo¹⁶; School of Human Evolution and Social Change, Arizona State University, Tempe,
29 USA, USA¹⁷; Institute of Vertebrate Biology, Academy of Sciences, Brno, Czech Republic¹⁸;
30 Department of Pathology and Parasitology, University of Veterinary and Pharmaceutical
31 Sciences, Brno, Czech Republic¹⁹; Biology Centre, Institute of Parasitology, Academy of
32 Sciences of the Czech Republic, Ceske Budejovice, Czech Republic²⁰; Liberec Zoo, Liberec,
33 Czech Republic²¹; World Wildlife Foundation (WWF), Dzanga Sangha Protected Areas,
34 Bangui, Central African Republic²²; Taï Chimpanzee Project, Centre Suisse de Recherches
35 Scientifiques, Abidjan, Côte d'Ivoire²³; Institute of Biology, University of Neuchatel,
36 Neuchatel, Switzerland²⁴; Budongo Conservation Field Station, Masindi, Uganda²⁵; School of
37 Psychology, University of St. Andrews, St. Andrews, Scotland, United Kingdom²⁶; Unité
38 Mixte Internationale 233, Institut de Recherche pour le Développement, INSERM U1175, and
39 University of Montpellier, Montpellier, France²⁷; Computational Biology Institute,
40 Montpellier, France²⁸

41

42 Running head: Great ape-polyomavirus co-divergence

43

44 # Address correspondence to Sébastien Calvignac-Spencer, calvignacs@rki.de.

45 § Present address: Department of Infection and Immunity, Luxembourg Institute of Health,
46 Esch-sur-Alzette, Luxembourg

47

48 Abstract: 133 words

49 Text: 4356 words

50

51 **Abstract**

52 It has long been hypothesized that polyomaviruses (PyV; family *Polyomaviridae*) co-diverged
53 with their animal hosts. In contrast, recent analyses suggested co-divergence may only
54 marginally influence the evolution of PyV. We re-assess this question by focusing on a single
55 lineage of PyV infecting hominine hosts, the Merkel cell polyomavirus (MCPyV) lineage. By
56 characterizing their genetic diversity in seven African great ape taxa, we show that these
57 viruses exhibit very strong host-specificity. Reconciliation analyses identify more co-
58 divergence than non co-divergence events. In addition, we find that a number of host and PyV
59 divergence events are synchronous. Collectively, our results support co-divergence as the
60 dominant process at play during the evolution of the MCPyV lineage. More generally, it adds
61 to the growing body of evidence suggesting an ancient and stable association of PyV and their
62 animal hosts.

63

64 **Importance**

65 The processes involved in viral evolution and the interaction of viruses with their hosts are of
66 great scientific interest and public health relevance. It has long been thought that the genetic
67 diversity of double stranded DNA viruses was generated over long periods of time, similar to
68 typical host evolutionary timescales. This was also hypothesized for polyomaviruses (family
69 *Polyomaviridae*), a group comprising several human pathogens, but this remains a point of
70 controversy. Here we investigate this question by focusing on a single lineage of
71 polyomaviruses that infect both humans and their closest relatives, the African great apes. We
72 show that these viruses exhibit considerable host-specificity and that their evolution largely
73 mirrors that of their hosts, suggesting co-divergence with their hosts played a major role in
74 their diversification. Our results provide statistical evidence in favor of an association of
75 polyomaviruses and their hosts over millions of years.

76

77 Introduction

78 Viral diversification is notably shaped by processes that promote host-specificity, e.g.
79 antagonistic co-evolution (1), and opportunities to colonize new hosts, i.e. cross-species
80 transmission events. Depending on their balance, host-virus co-divergence patterns may arise
81 and persist over the long term. Long-term co-divergence may have played an important role in
82 the diversification of some double stranded DNA (dsDNA) viruses, e.g. herpesviruses and
83 papillomaviruses (2-5).

84
85 Polyomaviruses (PyV; family *Polyomaviridae*) are small non-enveloped viruses with a
86 circular double stranded DNA genome (ca. 5 kb in length; 6). They infect a broad range of
87 animals, including arthropods and vertebrates (fish, birds and mammals), and comprise at least
88 13 distinct viruses infecting humans (7, 8). In humans, infections occur in childhood, persist
89 lifelong and are usually asymptomatic (9). At least 5 PyV have been associated with disease
90 in immunosuppressed individuals (10-12). Routes of transmission are poorly characterized but
91 may involve respiratory droplets and/or environmental contamination.

92
93 Putative co-divergence events of hosts and their PyV have repeatedly been evoked in the
94 literature to explain the structure of PyV diversity. Reconciliation analyses performed at the
95 family scale sometimes supported a significant contribution of co-divergence events (8, 13)
96 but others have failed to detect any global co-divergence signal (14, 15). Similarly, authors
97 focusing on more recent evolutionary events defended opposing views as to the potential co-
98 divergence of humans and JC polyomaviruses (JCV; 16, 17-19). An alternative scenario
99 combining ancient non co-divergence events and subsequent lineage-specific co-divergence
100 with their hosts, as proposed for papillomaviruses (3), still remains to be tested. The disparate
101 sampling of their animal hosts as well as the lack of resolution of many internal branches of

102 this viral family tree severely compromise the power to detect such patterns from currently
103 available data.
104
105 To overcome these limitations, we designed a formal test to assess the influence of co-
106 divergence on the evolution of PyV and characterized the genetic diversity of a single lineage
107 of PyV that infect a set of recently diverged host species with a well resolved phylogeny.
108 Specifically, we focused on viruses infecting African great apes (hereafter simply referred to
109 as great apes) belonging to the lineage comprising the Merkel cell polyomavirus, an
110 oncogenic human virus (MCPyV; *Human polyomavirus 5*, genus *Alphapolyomavirus*; 10, 20,
111 21, 22).

112

113 **Material and methods**

114

115 Samples

116 We collected a total of 386 fecal samples in the wild from seven great ape taxa (**Table 1**).
117 Great ape samples were collected opportunistically or from habituated animals, and preserved
118 in RNAlater (Qiagen, Hilden, Germany), in liquid nitrogen or by drying over silica. We also
119 collected 197 fecal samples from two human populations in Côte d'Ivoire and the Democratic
120 Republic of the Congo (**Table 1**). Human samples were preserved in liquid nitrogen. For
121 animal samples, authorization was obtained from responsible local authorities. For human
122 samples, institutional authorization was received along with the written consent of all
123 participants in the study.

124

125 Molecular biology

126 DNA extraction was performed using the Roboklon stool kit (Roboklon, Berlin, Germany),
127 according to manufacturer's instructions.

128

129 To identify Merkel cell polyomavirus-related (MCPyV-related) sequences in DNA extracts, a
130 nested PCR assay was set up that made use of generic, degenerate primers targeting a ca. 700
131 bp VP1 fragment (PCR1; **Table 2**). These primers were designed on the basis of published
132 MCPyV sequences and those of MCPyV-related PyV of nonhuman primates (NHP). First
133 round PCR mixes were set up so as to reduce the risk of carry over contamination with PCR
134 products. They contained 0.2 μ M of each primer, 200 μ M dNTP (with dUTP replacing
135 dTTP), 0.3 U AmpErase® uracil N-glycosylase (UNG; Invitrogen, Carlsbad, CA, USA), 4
136 mM MgCl₂, 1X PCR buffer and 1.25 U Platinum® Taq polymerase (Invitrogen). Second
137 round PCR mixes were prepared in the same way but did not include UNG. Cycling
138 conditions were as follows: 7 min at 45°C (UNG activity), 7 min at 95°C; 47 cycles (first
139 round) or 45 cycles (second round) of 30 s at 95°C, 30 s at 57°C (first round) or 58°C (second
140 round), and 2 min at 72°C; 10 min at 72°C.

141

142 Twenty-two positive samples were then selected based on the results of preliminary
143 phylogenetic analyses to attempt additional nested long-distance (LD) amplification of partial
144 genomes (approximately 2.5 kb) with generic, degenerate primers (PCR2; **Table 2**) using the
145 TaKaRa-Ex Kit (Takara Bio Inc., Otsu, Japan) according to manufacturer's instructions. Non-
146 degenerate primers (sequences available from the authors upon request) were used for
147 amplification of the remaining part (approximately 2.8 kb) of the genome with LD nested
148 PCR. LD PCR cycling conditions followed those reported in (21).

149

150 One hundred seventy-four human DNA extracts were also screened using a semi-nested PCR
151 system targeting a ca. 200bp VP1 fragment (PCR3; **Table 2**). This system was designed to be
152 specific to members of lineage 1 (see below) and was validated on a selection of great ape

153 DNA extracts of known status before being employed on human DNA extracts (data not
154 shown). PCR mix preparation and cycling conditions followed those mentioned above.

155

156 Short PCR products were purified using ExoSAP-IT (Affymetrix, Santa Clara, CA, USA)
157 whereas LD PCR products were purified using a column-based PCR purification kit (Qiagen,
158 Venlo, Netherlands). All purified products were sequenced with the Big Dye terminator cycle
159 sequencing kit on a 377 DNA automated sequencer (Applied Biosystems, Warrington, UK).

160

161 Overlapping partial sequences were used to reconstruct circular genomes using Geneious
162 v7.1.4 (Biomatters Ltd., Auckland, New Zealand; 23) . Genomes were subsequently annotated
163 with Geneious.

164

165 Phylogenetic analyses

166 Partial VP1 and complete genome datasets were assembled that comprised sequences
167 generated in this study and a selection (partial VP1) or all (complete genome) MCPyV
168 sequences as well as any publicly available great ape MCPyV-related sequence. Both datasets
169 were reduced to unique sequences and aligned using MUSCLE, as implemented in SeaView
170 v4 (24). Conserved nucleotide blocks were selected from the alignments using Gblocks (still
171 in SeaView; 25) and used for recombination analyses using RDP4 v4.46 (26). The final
172 alignments comprised 74 sequences and 838 positions (partial VP1) and 16 sequences and
173 5150 positions (complete genome). Further analyses were performed only on the partial VP1
174 alignment, as this comprised the most genetic diversity in this dataset.

175

176 The best model of nucleotide substitution (general time reversible matrix with rate variation
177 across sites; GTR+G₄) was selected with jModelTest v2.1.4 (27), using the Bayesian
178 information criterion. Maximum likelihood analyses were performed using PhyML v3 (28), as

179 implemented on the PhyML webserver (29). The ‘best-fit’ root of the ML tree was identified
180 using TempEst v1.5 (<http://tree.bio.ed.ac.uk/software/tempest/>; 30). Bayesian Markov chain
181 Monte Carlo (BMCMC) analyses were performed in BEAST v1.8.2 under a lognormal
182 relaxed clock (uncorrelated) and three different models of diversification: a pure coalescent
183 model assuming a constant population size, a multi-species coalescent model using the 14-
184 species scheme suggested by species delineation analyses (see below), and a birth-death
185 speciation model (31, 32). Convergence of BMCMC runs (at least two runs per model) and
186 appropriate sampling of the posterior were checked with Tracer v1.6
187 (<http://tree.bio.ed.ac.uk/software/tracer/>). Branch robustness was assessed through non-
188 parametric bootstrapping (250 pseudo-replicates; ML) or posterior probabilities (BMCMC).

189

190 Host specificity analyses

191 Host specificity was assessed by running BaTS on all posterior samples of trees (PST)
192 generated by BMCMC analyses (33). BaTS allows for tests of the correlation of trait states
193 with ancestry while accounting for phylogenetic uncertainty suggested by the PST. It
194 compares observations to a null distribution generated under the assumption that trait values
195 are not influenced by ancestry. Host species/sub-species was defined as the trait of interest. Its
196 association with ancestry was assessed at the host sub-species level (8 states) and species
197 level (5 states) independently, by running separate BaTS analyses during which 500 null
198 replicates per tree were generated. Global as well as state-specific statistics of association
199 were computed (global: association index, AI, and Fitch parsimony score, PS; state-specific:
200 maximum exclusive single-state clade size, MC)

201

202 To investigate the association of host and PyV diversification processes, we performed PyV
203 species delineation analyses with the R package *splits* (34), using the maximum clade
204 credibility tree derived from BMCMC analyses performed under the (coalescent) constant

205 population size model. *Splits* implements general mixed Yule-coalescent models (GMYC; 34,
206 35) which are optimized and compared to the null hypothesis that the tree was generated by
207 pure coalescent processes, i.e. reflects diversity within a single species. When the GMYC
208 model outperforms the null model, the parts of the tree most likely to have been generated by
209 between-species and within-species processes can be identified, thereby delineating species
210 (according to the phylogenetic species concept).

211

212 Co-divergence analyses

213 The degree of topological congruence and the number of events necessary to explain
214 (reconcile) incongruences were assessed using Jane v4 (36). Jane implements a genetic
215 algorithm to quickly identify the most parsimonious scenarios of co-evolution, involving
216 several types of events (co-divergence, duplication, duplication with host switch, loss and
217 failure to diverge). As input, it requires host and parasite phylogenies and the according tip
218 mapping as well as an event cost matrix. A simplified version of the PyV phylogeny was used
219 as input, whereby single-host clades were collapsed. Three sets of costs were tested: i) set 1:
220 co-divergence 0, duplication 1 (under the assumption that duplication incurs costs related to
221 within-host speciation, e.g. maintaining of distinct lineages in the face of within-host
222 competition or tropism change within the same host), duplication with host switch 1 (host
223 switch incurs costs), loss 1 (prevalence was always high) and failure to diverge 1 (given their
224 respective evolutionary timescales, viruses are unlikely to fail to diverge when their hosts do
225 so), ii) set 2: same as set 1 but with loss 0 (prevalence may have been low at some point in the
226 past), iii) set 3: co-divergence -1, all non co-divergence events 0. Set 3 is a variation of set 1
227 with the same relative costs but where all costs are shifted to the left. This allows equating
228 costs and co-divergence events. Jane was run using the vertex-based cost mode and the
229 parameters of the genetic algorithm were kept at their default values (population size 100,
230 number of generations 100). To determine the probability of observing the inferred costs by

231 chance, costs were also calculated on a set of 500 samples for which tip mapping was
232 randomized. Settings of the genetic algorithm were kept at default values.

233

234 Topology tests were performed to assess whether exceptions to a scenario of perfect co-
235 divergence observed in the PyV phylogenetic tree were better supported by the data than a
236 perfect co-divergence model. This was done by using approximately unbiased tests (AU-
237 tests), as implemented in CONSEL v0.1i (37).

238

239 Finally, divergence dates were also estimated. Topological congruence could emerge
240 independently of co-divergence, e.g. through preferential host switching (38). Observing
241 synchronicity in timing of divergence events of hosts and their parasites reinforce the co-
242 divergence hypothesis. When viral lineage duplication occurs, synchronicity of parasite
243 divergence events is also expected (provided the viral lineages maintain similar degrees of
244 association to their host). Divergence date estimates were obtained using two methods: i) as
245 part of the aforementioned BMCMC analyses, or ii) by re-estimating branch lengths of the
246 ML tree under codon models using HyPhy v2.2.4 (39) and making the resulting tree
247 ultrametric using a relaxed clock model implemented in r8s (40). The codon models used for
248 this second set of analyses were a pure branch model derived of MG94 in which the ratio of
249 nonsynonymous substitutions per nonsynonymous site to synonymous substitutions per
250 synonymous site is estimated for each branch but assumed to be unchanged across sites (41)
251 and an adaptive branch-site random effects model in which this ratio is estimated for each
252 branch and allowed to vary across sites (aBSREL; 42). We detected marked saturation at
253 synonymous sites (data not shown); such strong saturation complicates analyses under both
254 nucleotide and codon models. For both BMCMC and ML-based analyses, the relaxed clock
255 was calibrated by setting a prior distribution (BEAST) or enforcing a fixed age (r8s) for the
256 time to the most recent common ancestor of lineage 1 using a published estimate of the split

257 date of all hominine species (either 5.6 My or a normal distribution of mean 5.6 My and
258 standard deviation 0.25 My; 43). Because we used the split date of all hominine species,
259 estimates of times to the most recent common ancestors for viruses should be regarded as
260 minimum bounds (viral coalescence times will necessarily predate the effective ancestral host
261 population/species split). It should also be noted that divergence dates of the different
262 hominine lineages are a point of active debated; this stems from both a scarce paleontological
263 record and uncertainty in estimates of long-term mutation rates at genomic scales. For
264 example, the estimate we opted for here (5.6 My) is drawn from genomic analyses that
265 proposed two estimates (5.6 or 11.2 My), depending on priors on the substitution rates (1 or
266 0.5×10^{-9} mutation. $\text{bp}^{-1} \cdot \text{year}^{-1}$; 43). The focus of our synchronicity analyses was, however, on
267 relative internode lengths, not absolute dates. Calendar years can thus be replaced with
268 genetic distances and/or ratios of interest (see **Table 6**).

269

270 **Results**

271

272 Detection of short MCPyV-related sequences

273 Using a specific PCR system designed to amplify a ca. 700 bp fragment of the VP1 gene, we
274 screened 386 fecal great ape and 197 human samples (**Table 1**). We detected MCPyV-related
275 sequences in 50 great ape DNA extracts representing all hosts but *G. g. diehli*, with fecal
276 detection rates between 1.2% (*G. b. beringei*) and 53.8% (*P. paniscus*). Nearly all sequences
277 were only found at one site; a single sequence was detected in 5 and 2 Eastern chimpanzees
278 (*P. t. schweinfurthii*) at two distinct sites in Uganda. For species/sub-species from which more
279 than 2 sequences were obtained, considerable sequence divergence was observed, e.g.
280 maximum observed distances were over 20%, possibly reflecting the circulation of viruses
281 belonging to different lineages (discussed in more detail below). Minimum observed distances
282 to publicly available sequences were often relatively high, i.e. between 5 and 17%. Finally,

283 we also detected MCPyV sequences – with > 99% identity to published MCPyV sequences -
284 in 30 human DNA extracts (fecal detection rate: 15.2%). Most human DNA extracts were also
285 screened with a PCR system intended to be lineage 1-specific (see below); all assays were
286 negative.

287

288 Characterization of full genomes

289 We attempted to determine full genome sequences from a selection of DNA extracts (N=22).
290 This was possible for samples from *P. paniscus* (N=2), *P. t. troglodytes* (N=3), *P. t.*
291 *schweinfurthii* (N=1) and *G. b. graueri* (N=1). Examination of putative open reading frames
292 (ORFs) showed that all genomes displayed a typical PyV genome structure with an early
293 region encoding regulatory proteins (small t and large T antigens) and a late region coding for
294 structural proteins (VP1, VP2 and VP3) separated by a non-coding control region (NCCR).
295 No open reading frame likely to encode a putative agnoprotein was identified. Overall, a ca.
296 80% sequence similarity to genomes of MCPyV and MCPyV-related nonhuman primate PyV
297 was observed. Preliminary analyses revealed that the full genomes represented only a fraction
298 of the overall genetic diversity detected in this study. To incorporate this broader diversity, we
299 performed all following phylogenetic analyses on an alignment of partial VP1 sequences
300 (including sequences extracted from the novel full genomes).

301

302 Molecular phylogeny

303 We could not detect any signal indicative of recombination in the VP1 alignment (26).
304 Phylogenetic analyses in both maximum likelihood (ML; 28) and Bayesian (31) frameworks
305 supported the existence of a number of host-specific clades (**Figure 1** and **Figure 2**). All
306 clades seemed to derive from three ancient lineages: one that only comprised MCPyV
307 sequences, and two that only included viral sequences detected in gorillas, bonobos and
308 chimpanzees. Branching order partially recapitulated host divergence events in the two great

ape lineages (hereafter referred to as lineages 1 and 2; **Figure 1** and **Figure 2**). We identified four exceptions: i) the polyphylies of PyV infecting Western chimpanzees in lineage 1 and Eastern chimpanzees in lineage 2, ii) the interspersions of PyV infecting Eastern lowland and mountain gorillas in lineage 1, iii) the basal position of MCPyV.

313

314 Host specificity

315 We estimated the statistical support for host specificity using BaTS (**Table 3**). We found that 316 viral sequences found in a single host species were generally more likely to be closely related 317 than expected by chance, when considering both global and state-specific statistics. The only 318 exceptions corresponded to viral sequences identified in the sister sub-species *G. b. beringei* 319 and *G. b. graueri*.

320

321 We also characterized the viral diversification process by running a species delineation 322 analysis using general mixed Yule-coalescent models (GMYC; 34, 35). The best GMYC 323 model outperformed the null, full coalescent model ($P=0.0005$) and identified 14 entities, 324 among which 10 comprised several sequences. Nine multi-sequence entities only comprised 325 sequences identified from a single host species/sub-species, indicating a close parallelism of 326 PyV and host diversification processes (**Figure 1**).

327

328 Co-divergence

329 Taking the viral phylogeny presented in **Figure 1** as a given, we performed reconciliation 330 analyses using Jane (**Table 4**). Under all tested cost sets, and whether the host species or sub- 331 species phylogeny was considered, the number of co-divergence events always exceeded the 332 number of non co-divergence events. Randomization tests showed that, irrespective of the 333 cost set, these results could not be explained by chance at the sub-species level. At the species 334 level and using a p-value threshold of 0.05, results obtained under two of the cost sets failed

335 to reach statistical significance; it should however be noted that the species-level phylogeny
336 only comprises 5 species, meaning these tests had low power.

337

338 We also examined whether the viral topology presented in **Figure 1** was a better fit to our
339 data than alternative topologies which enforced strict co-divergence within lineages 1 and 2.
340 The model forcing MCPyV to belong to lineage 1 was the only that was rejected (AU-test;
341 $P=0.003$). Monophyly of PyV infecting Western chimpanzees in lineage 1 and Eastern
342 chimpanzees in lineage 2 as well as inclusion of MCPyV in lineage 2 could not be excluded
343 (AU-test; $P=0.52$, 0.13 and 0.11). Given the very recent split of Eastern lowland and
344 mountain gorillas (about 10000 years ago; 44), the interspersions of PyV infecting these
345 subspecies appeared biologically plausible, so we did not compare this scenario to a strict co-
346 divergence model.

347

348 Besides topological congruence, co-divergence should result in synchronization of: i) viral
349 and host divergence dates and ii) viral divergence dates in the case of ancestral viral lineage
350 duplication. We first estimated divergence dates using a relaxed clock model applied to
351 nucleotide data in a Bayesian framework. For 5 of the 6 focal nodes of our analyses (nodes
352 1.2 to 4 and 2.1 to 3), these estimates were significantly older than host divergence events
353 (**Table 5**). This pattern was compatible with the effects of the time dependency of molecular
354 rates – i.e. the decay of molecular rates with increasing observation timescales - which can
355 result in overestimating recent time to the most recent common ancestor (tMRCA) inferred
356 from deep calibration points (19, 45-47). As this may arise through the effects of
357 unaccounted-for purifying selection (amongst other possible mechanisms; 48, 49, 50), we re-
358 estimated all branch lengths using selection-aware models of codon evolution in a ML
359 framework. A branch model of codon evolution resulted in divergence dates very close to
360 those inferred by BMCMC analyses. Using an adaptive branch-site random effect model of

361 codon evolution, strong purifying selection was detected on a number of branches, including
362 deep ones (data not shown). Most of the resulting increase in the overall tree length was
363 supported by a single basal branch. This expansion prevented deriving any trustworthy
364 tMRCA estimates.

365
366 Given the likely impact of strong purifying selection and our inability to properly account for
367 it, we re-examined branch length/internode ratios by re-scaling the results in **Table 5**, using
368 the tMRCA of a young node –node 1.4 (divergence of lineage 1 PyV infecting *P. t.*
369 *trogodytes* and *P. t. schweinfurthii*) - as a new arbitrary unit (**Table 6**). This resulted in a
370 good agreement of host and virus relative divergence dates for most nodes (nodes 1.3 and 2.3
371 and nodes 1.2 and 2.2). The tMRCA of lineage 2 PyV infecting all great apes was a large
372 underestimate of the divergence date of their hominine hosts, as expected under the
373 hypothesis that deep branch lengths are severely underestimated.

374 375 **Discussion**

376 The lack of any physical viral fossil record considerably complicates the task of
377 understanding the long-term association of viruses with their hosts. However, using their
378 present-day distribution, their nucleic acid sequences and (more rarely) other biological traits,
379 we can try to infer how long and how closely viruses have been associated to their hosts. The
380 aim of this study was to determine whether co-divergence, i.e. viral diversification driven by
381 host diversification, is an important driver of PyV evolution.

382
383 Measurable host specificity is an absolute prerequisite for characterizing historical co-
384 divergence events. Host specificity has often been assumed for PyV, with only a few well-
385 identified exceptions, e.g. budgerigar fledgling disease virus and SV40. Over the last decade,
386 this assumption has been repeatedly supported by the implementation of generic PyV

387 detection tools which have not revealed any multi-host PyV (20, 51). Here, we used a PCR
388 assay designed to specifically target a single PyV lineage to generate a large sample of
389 sequences from closely related PyV infecting wild African great apes. Statistical tests strongly
390 supported marked host specificity, which was still detectable at the host sub-species level.

391 Viral diversification/speciation - as revealed by a GMYC model, i.e. according to the
392 phylogenetic species concept - appeared strongly influenced by host diversification.

393

394 Host specificity and a coupling of viral diversification/speciation with host diversity could
395 also arise over much shorter timescales than those implied by co-divergence events. If co-
396 divergence is a dominant evolutionary process a key expectation is that virus and host
397 phylogenies should often be congruent. Phylogenetic analyses of great ape MCPyV-like
398 sequences highlighted the existence of two viral lineages within which viral divergence events
399 were mostly in line with hominine divergence events. Exceptions to the expectation of perfect
400 co-divergence within these lineages were not statistically supported. In addition,
401 reconciliation analyses identified more co-divergence events than non co-divergence events,
402 irrespective of the host taxonomic level and cost set, e.g. 10 co-divergence events vs. 5 non
403 co-divergence events considering host sub-species and all cost sets. Co-divergence may
404 therefore be the dominant process at play, accompanied by less frequent non co-divergence
405 events, e.g. the viral lineage duplication event that gave rise to lineages 1 and 2.

406

407 On short timescales, host relatedness may influence viral transmission in such a way that
408 topological congruence ensues in the absence of real co-divergence, e.g. if host jumps are
409 facilitated by host phylogenetic proximity (the preferential host switch hypothesis; 38, 52). A
410 further step in validating co-divergence events consists of showing that host and virus
411 divergence events are synchronized. This requires branch lengths to be properly estimated
412 throughout the phylogeny. Here, we speculate that the well-documented time dependency of

413 molecular rates – which posits an apparent decay of molecular rates with increasing
414 measurement timescales (19, 45-47) - may have resulted in overestimating recent divergence
415 dates derived from our initial molecular clock analyses which were calibrated with an ancient
416 divergence event. In line with this hypothesis, we found that the relative timescales of host
417 and virus divergence events were in good agreement when these estimates were re-scaled
418 using an arbitrary unit set to a recent divergence event, i.e. a procedure similar to calibrating
419 the molecular clock with this recent divergence event. In addition, co-divergence events were
420 also synchronous in the viral lineages 1 and 2.

421

422 Overall, we observe i) marked host-specificity, ii) frequent co-divergence events and iii) the
423 synchronicity of a number of co-divergence events. The evolution of MCPyV-related viruses
424 with their hominine hosts therefore appears to have been mostly driven by host-PyV co-
425 divergence. A number of other human PyV have been shown to be closely related to great ape
426 PyV (22, 53-56). The according lineages may represent promising opportunities to test
427 whether the dominance of co-divergence events can be generalized throughout the PyV
428 family tree. Regardless, the findings reported here lend support to the hypothesis of an ancient
429 association of PyV and their animal hosts, which the well-known separation of mammal and
430 bird PyV and the recent discovery of the first fish and arthropod PyV already pinpointed (6, 8,
431 57). In a recently published LT phylogeny, the root age of the family tree was more than 11
432 times the age of the MRCA of MCPyV-related viruses (20). Assuming this MRCA dates back
433 to about 6 My ago, the family root would be more than 60 My old. Assuming that the PyV
434 family tree is affected by the phenomenon of time dependency of molecular rates, the root age
435 of the family may be even more ancient, as recently suggested by C. B. Buck, et al. (8).

436

437 Although a robust signal for co-divergence exists, we did not observe strict co-divergence of
438 MCPyV-related viruses and their hominine hosts. For example, in our phylogenetic analyses

the placement of the MCPyV lineage is ambiguous and the most ancient divergence event of polyomaviruses apparently post-dates the according divergence event of their hominine hosts. Although these observations may be explained by limitations of the models of sequence evolution we used, we cannot exclude the hypothesis that they reflect biological reality. Since hominine species are recently diverged, the combination of ancestral viral diversity and incomplete lineage sorting may suffice to explain apparent deviations from strict co-divergence, i.e. perfect patterns of co-divergence are not necessarily expected, even where no other processes have been at play (19). However, a notion emerging in the literature is that a mixture of processes, including but not restricted to measurable co-divergence with their hosts, will generally provide a better explanation for dsDNA virus evolution in the long run than strict co-divergence. For example, it was proposed that HSV-2 arose as a consequence of the transmission of a chimpanzee simplexvirus to the human lineage (50). Similarly, host switches as well as lineage duplications have been documented in papillomaviruses (2, 3). It seems clear that processes other than co-divergence were also at play during PyV evolution, as notably illustrated by the 13 human PyV identified thus far and the two great ape lineages documented in this study. Further biological characterization of representatives of these lineages may reveal whether these non co-divergence events were driven by adaptive, e.g. tissue tropism change, or stochastic, e.g. demographic, processes (58).

457

458 **Funding information**

Research on Central chimpanzees, Cross-River gorillas and Western lowland gorillas in Cameroon received support from Working Dogs for Conservation, the Wildlife Conservation Society (WCS), the United States Fish and Wildlife Service Great Ape Conservation Fund, the Association of Zoos and Aquariums Conservation Endowment Fund and the Agence Nationale de Recherches sur le SIDA (ANRS), France (ANRS 12125, ANRS 12182, ANRS 12555, and ANRS 12325). Research on Western lowland gorillas in the Central African

465 Republic received support from the Primate Habituation Programme, the World Wildlife
466 Fund, the Grant Agency of the Czech Republic (#206/09/0927), the Institute of Vertebrate
467 Biology of the Academy of Sciences of the Czech Republic (#RVO68081766), the European
468 Social Fund and a Praemium Academiae award to J. Lukes. Research on Central chimpanzees
469 and Western lowland gorillas in Gabon received support from WCS and the Société pour la
470 Conservation et le Développement. Research on Eastern chimpanzees in Uganda received
471 support from the Royal Zoological Society of Scotland (core funding of the Budongo
472 Conservation Field Station; BCFS). Research on Eastern lowland and mountain gorillas in the
473 Democratic Republic of the Congo, Rwanda and Uganda received support from WWF
474 Sweden, the Fair Play Foundation, the Netherlands Directorate General for International
475 Cooperation through the Greater Virunga Transboundary Collaboration, the Berggorilla &
476 Regenwald Direkthilfe e.V., the Max Planck Society and WCS. JOW was funded by the NIH
477 (K01-AI110181) and the University of California Laboratory Fees Research Program (grant
478 number: 12-LR-236617). The funders had no role in study design, data collection and
479 interpretation, or the decision to submit the work for publication.

480

481 **Acknowledgements**

482 Research on Central chimpanzees, Cross-River gorillas and Western lowland gorillas in
483 Cameroon was conducted with permission from the Ministries of Health, Forests and Wildlife
484 and Research and benefited from the assistance R. Ikfuingei in the field and D. Ryu in the lab.
485 Research on Western lowland gorillas in the Central African Republic was conducted with
486 permission from the Ministère de l'Éducation Nationale, de l'Alphabétisation, de
487 l'Enseignement Supérieur et de la Recherche and benefited from the assistance of the staff of
488 Dzanga-Sangha Protected Areas and local trackers and assistants. Research on bonobos in the
489 Democratic Republic of the Congo (DRC) was conducted with permission of the Institut
490 Congolais pour la Conservation de la Nature (ICCN) and benefited from the assistance of

491 students and field assistants of the Salonga Bonobo Project. Research on Central chimpanzees
492 and Western lowland gorillas in Gabon was conducted with permission from the Agence
493 Nationale des Parcs Nationaux (ANPN) and the Centre de la Recherche Scientifique et
494 Technologique and benefited from the assistance of WCS and ANPN staff as well as field
495 assistants of the Loango Ape Project. Research on Eastern chimpanzees in Uganda was
496 conducted with permission from the Uganda Wildlife Authority, the Uganda National Council
497 for Science and Technology and the Makerere University Biological Field Station (MUBFS)
498 and benefited from the assistance of many students and assistants at the BCFS and MUBFS.
499 Research on Eastern lowland and mountain gorillas in DRC, Rwanda and Uganda was
500 conducted with permission from ICCN, the Rwandan Development Board and UWA and
501 benefited from the assistance of the Mountain Gorilla Veterinary Project, the Dian Fossey
502 Gorilla Fund International and Conservation Through Public Health for the organization of
503 the censuses. We are grateful to J. Gogarten for his careful proofreading of this manuscript.
504

505 **Data availability**

506 Partial VP1 and whole genome sequences were deposited at the European Nucleotide Archive
507 and GenBank, respectively, under accession numbers LT158307-LT158400 and KT184856-
508 KT184862. r8s and BEAUTi XML exemplary input files are available from the authors upon
509 request.
510

511 **References**

- 512 1. **Compton AA, Emerman M.** 2013. Convergence and divergence in the evolution of
513 the APOBEC3G-Vif interaction reveal ancient origins of simian immunodeficiency
514 viruses. *PLoS Pathog* **9**:e1003135.
- 515 2. **Garcia-Perez R, Ibanez C, Godinez JM, Arechiga N, Garin I, Perez-Suarez G, de**
516 **Paz O, Juste J, Echevarria JE, Bravo IG.** 2014. Novel papillomaviruses in free-
517 ranging Iberian bats: no virus-host co-evolution, no strict host specificity, and hints for
518 recombination. *Genome Biol Evol* **6**:94-104.

- 519 3. **Gottschling M, Goker M, Stamatakis A, Bininda-Emonds OR, Nindl I, Bravo IG.**
520 2011. Quantifying the phylodynamic forces driving papillomavirus evolution. *Mol*
521 *Biol Evol* **28**:2101-2113.
- 522 4. **Lavergne A, Donato D, Gessain A, Niphuis H, Nerrienet E, Verschoor EJ,**
523 **Lacoste V.** 2014. African great apes are naturally infected with roseoloviruses closely
524 related to human herpesvirus 7. *J Virol* **88**:13212-13220.
- 525 5. **McGeoch DJ, Rixon FJ, Davison AJ.** 2006. Topics in herpesvirus genomics and
526 evolution. *Virus Res* **117**:90-104.
- 527 6. **Johne R, Buck CB, Allander T, Atwood WJ, Garcea RL, Imperiale MJ, Major**
528 **EO, Ramqvist T, Norkin LC.** 2011. Taxonomical developments in the family
529 Polyomaviridae. *Arch Virol* **156**:1627-1634.
- 530 7. **Mishra N, Pereira M, Rhodes RH, An P, Pipas JM, Jain K, Kapoor A, Briese T,**
531 **Faust PL, Lipkin WI.** 2014. Identification of a novel polyomavirus in a pancreatic
532 transplant recipient with retinal blindness and vasculitic myopathy. *J Infect Dis*
533 **210**:1595-1599.
- 534 8. **Buck CB, Van Doorslaer K, Peretti A, Geoghegan EM, Tisza MJ, An P, Katz JP,**
535 **Pipas JM, McBride AA, Camus AC, McDermott AJ, Dill JA, Delwart E, Ng TF,**
536 **Farkas K, Austin C, Kraberger S, Davison W, Pastrana DV, Varsani A.** 2016.
537 The Ancient Evolutionary History of Polyomaviruses. *PLoS Pathog* **12**:e1005574.
- 538 9. **Rockett RJ, Bialasiewicz S, Mhango L, Gaydon J, Holding R, Whiley DM,**
539 **Lambert SB, Ware RS, Nissen MD, Grimwood K, Sloots TP.** 2015. Acquisition of
540 human polyomaviruses in the first 18 months of life. *Emerg Infect Dis* **21**:365-367.
- 541 10. **Feng H, Shuda M, Chang Y, Moore PS.** 2008. Clonal integration of a polyomavirus
542 in human Merkel cell carcinoma. *Science* **319**:1096-1100.
- 543 11. **Ho J, Jedrych JJ, Feng H, Natalie AA, Grandinetti L, Mirvish E, Crespo MM,**
544 **Yadav D, Fasanella KE, Proksell S, Kuan SF, Pastrana DV, Buck CB, Shuda Y,**
545 **Moore PS, Chang Y.** 2015. Human polyomavirus 7-associated pruritic rash and
546 viremia in transplant recipients. *J Infect Dis* **211**:1560-1565.
- 547 12. **van der Meijden E, Janssens RW, Lauber C, Bouwes Bavinck JN, Gorbalenya**
548 **AE, Feltkamp MC.** 2010. Discovery of a new human polyomavirus associated with
549 trichodysplasia spinulosa in an immunocompromized patient. *PLoS Pathog*
550 **6**:e1001024.
- 551 13. **Perez-Losada M, Christensen RG, McClellan DA, Adams BJ, Viscidi RP,**
552 **Demma JC, Crandall KA.** 2006. Comparing phylogenetic codivergence between
553 polyomaviruses and their hosts. *J Virol* **80**:5663-5669.
- 554 14. **Krumbholz A, Bininda-Emonds OR, Wutzler P, Zell R.** 2009. Phylogenetics,
555 evolution, and medical importance of polyomaviruses. *Infect Genet Evol* **9**:784-799.
- 556 15. **Tao Y, Shi M, Conrardy C, Kuzmin IV, Recuenco S, Agwanda B, Alvarez DA,**
557 **Ellison JA, Gilbert AT, Moran D, Niezgoda M, Lindblade KA, Holmes EC,**

- 558 **Breiman RF, Rupprecht CE, Tong S.** 2013. Discovery of diverse polyomaviruses in
559 bats and the evolutionary history of the Polyomaviridae. *J Gen Virol* **94**:738-748.
- 560 16. **Agostini HT, Yanagihara R, Davis V, Ryschkewitsch CF, Stoner GL.** 1997. Asian
561 genotypes of JC virus in Native Americans and in a Pacific Island population: markers
562 of viral evolution and human migration. *Proc Natl Acad Sci U S A* **94**:14542-14546.
- 563 17. **Shackelton LA, Rambaut A, Pybus OG, Holmes EC.** 2006. JC virus evolution and
564 its association with human populations. *J Virol* **80**:9928-9933.
- 565 18. **Sugimoto C, Kitamura T, Guo J, Al-Ahdal MN, Shchelkunov SN, Otova B,**
566 **Ondrejka P, Chollet JY, El-Safi S, Ettayebi M, Gresenguet G, Kocagoz T,**
567 **Chaiyarasamee S, Thant KZ, Thein S, Moe K, Kobayashi N, Taguchi F, Yogo Y.**
568 1997. Typing of urinary JC virus DNA offers a novel means of tracing human
569 migrations. *Proc Natl Acad Sci U S A* **94**:9191-9196.
- 570 19. **Sharp PM, Simmonds P.** 2011. Evaluating the evidence for virus/host co-evolution.
571 *Curr Opin Virol* **1**:436-441.
- 572 20. **Calvignac-Spencer S, Feltkamp MC, Daugherty MD, Moens U, Ramqvist T,**
573 **Johne R, Ehlers B.** 2016. A taxonomy update for the family *Polyomaviridae*. *Arch*
574 *Virol*.
- 575 21. **Leendertz FH, Scuda N, Cameron KN, Kidega T, Zuberbuhler K, Leendertz SA,**
576 **Couacy-Hymann E, Boesch C, Calvignac S, Ehlers B.** 2011. African great apes are
577 naturally infected with polyomaviruses closely related to Merkel cell polyomavirus. *J*
578 *Virol* **85**:916-924.
- 579 22. **Scuda N, Madinda NF, Akoua-Koffi C, Adjogoua EV, Wevers D, Hofmann J,**
580 **Cameron KN, Leendertz SA, Couacy-Hymann E, Robbins M, Boesch C, Jarvis**
581 **MA, Moens U, Mugisha L, Calvignac-Spencer S, Leendertz FH, Ehlers B.** 2013.
582 Novel polyomaviruses of nonhuman primates: genetic and serological predictors for
583 the existence of multiple unknown polyomaviruses within the human population.
584 *PLoS Pathog* **9**:e1003429.
- 585 23. **Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S,**
586 **Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond**
587 **A.** 2012. Geneious Basic: an integrated and extendable desktop software platform for
588 the organization and analysis of sequence data. *Bioinformatics* **28**:1647-1649.
- 589 24. **Gouy M, Guindon S, Gascuel O.** 2010. SeaView version 4: A multiplatform
590 graphical user interface for sequence alignment and phylogenetic tree building. *Mol*
591 *Biol Evol* **27**:221-224.
- 592 25. **Talavera G, Castresana J.** 2007. Improvement of phylogenies after removing
593 divergent and ambiguously aligned blocks from protein sequence alignments. *Syst*
594 *Biol* **56**:564-577.
- 595 26. **Martin DP, Murrell B, Golden M, Khoosal A, Muhire B.** 2015. RDP4: Detection
596 and analysis of recombination patterns in virus genomes. *Virus Evolution* **1**.

- 597 27. **Darriba D, Taboada GL, Doallo R, Posada D.** 2012. jModelTest 2: more models,
598 new heuristics and parallel computing. *Nat Methods* **9**:772.
- 599 28. **Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O.** 2010.
600 New algorithms and methods to estimate maximum-likelihood phylogenies: assessing
601 the performance of PhyML 3.0. *Syst Biol* **59**:307-321.
- 602 29. **Guindon S, Lethiec F, Duroux P, Gascuel O.** 2005. PHYML Online--a web server
603 for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Res*
604 **33**:W557-559.
- 605 30. **Rambaut A, Lam TT, Max Carvalho L, Pybus OG.** 2016. Exploring the temporal
606 structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus*
607 *Evolution* **2**.
- 608 31. **Drummond AJ, Suchard MA, Xie D, Rambaut A.** 2012. Bayesian phylogenetics
609 with BEAUti and the BEAST 1.7. *Mol Biol Evol* **29**:1969-1973.
- 610 32. **Heled J, Drummond AJ.** 2010. Bayesian inference of species trees from multilocus
611 data. *Mol Biol Evol* **27**:570-580.
- 612 33. **Parker J, Rambaut A, Pybus OG.** 2008. Correlating viral phenotypes with
613 phylogeny: accounting for phylogenetic uncertainty. *Infect Genet Evol* **8**:239-246.
- 614 34. **Fujisawa T, Barraclough TG.** 2013. Delimiting species using single-locus data and
615 the Generalized Mixed Yule Coalescent approach: a revised method and evaluation on
616 simulated data sets. *Syst Biol* **62**:707-724.
- 617 35. **Pons J, Barraclough TG, Gomez-Zurita J, Cardoso A, Duran DP, Hazell S,**
618 **Kamoun S, Sumlin WD, Vogler AP.** 2006. Sequence-based species delimitation for
619 the DNA taxonomy of undescribed insects. *Syst Biol* **55**:595-609.
- 620 36. **Conow C, Fielder D, Ovadia Y, Libeskind-Hadas R.** 2010. Jane: a new tool for the
621 cophylogeny reconstruction problem. *Algorithms Mol Biol* **5**:16.
- 622 37. **Shimodaira H, Hasegawa M.** 2001. CONSEL: for assessing the confidence of
623 phylogenetic tree selection. *Bioinformatics* **17**:1246-1247.
- 624 38. **Charleston MA, Robertson DL.** 2002. Preferential host switching by primate
625 lentiviruses can account for phylogenetic similarity with the primate phylogeny. *Syst*
626 *Biol* **51**:528-535.
- 627 39. **Pond SL, Frost SD, Muse SV.** 2005. HyPhy: hypothesis testing using phylogenies.
628 *Bioinformatics* **21**:676-679.
- 629 40. **Sanderson MJ.** 2003. r8s: inferring absolute rates of molecular evolution and
630 divergence times in the absence of a molecular clock. *Bioinformatics* **19**:301-302.
- 631 41. **Muse SV, Gaut BS.** 1994. A likelihood approach for comparing synonymous and
632 nonsynonymous nucleotide substitution rates, with application to the chloroplast
633 genome. *Mol Biol Evol* **11**:715-724.

- 634 42. **Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Kosakovsky Pond**
635 **SL.** 2015. Less is more: an adaptive branch-site random effects model for efficient
636 detection of episodic diversifying selection. *Mol Biol Evol* **32**:1342-1353.
- 637 43. **Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B,**
638 **Veeramah KR, Woerner AE, O'Connor TD, Santpere G, Cagan A, Theunert C,**
639 **Casals F, Laayouni H, Munch K, Hobolth A, Halager AE, Malig M, Hernandez-**
640 **Rodriguez J, Hernando-Herraez I, Prufer K, Pybus M, Johnstone L, Lachmann**
641 **M, Alkan C, Twigg D, Petit N, Baker C, Hormozdiari F, Fernandez-Callejo M,**
642 **Dabad M, Wilson ML, Stevison L, Camprubi C, Carvalho T, Ruiz-Herrera A,**
643 **Vives L, Mele M, Abello T, Kondova I, Bontrop RE, Pusey A, Lankester F,**
644 **Kiyang JA, Bergl RA, Lonsdorf E, Myers S, Ventura M, Gagneux P, Comas D, et**
645 **al.** 2013. Great ape genetic diversity and population history. *Nature* **499**:471-475.
- 646 44. **Roy J, Arandjelovic M, Bradley BJ, Guschanski K, Stephens CR, Bucknell D,**
647 **Cirhuza H, Kusamba C, Kyungu JC, Smith V, Robbins MM, Vigilant L.** 2014.
648 Recent divergences and size decreases of eastern gorilla populations. *Biol Lett*
649 **10**:20140811.
- 650 45. **Aiewsakun P, Katzourakis A.** 2016. Time-dependent rate phenomenon in viruses. *J*
651 *Virol* doi:10.1128/jvi.00593-16.
- 652 46. **Duchene S, Holmes EC, Ho SY.** 2014. Analyses of evolutionary dynamics in viruses
653 are hindered by a time-dependent bias in rate estimates. *Proc Biol Sci* **281**.
- 654 47. **Ho SY, Phillips MJ, Cooper A, Drummond AJ.** 2005. Time dependency of
655 molecular rate estimates and systematic overestimation of recent divergence times.
656 *Mol Biol Evol* **22**:1561-1568.
- 657 48. **Wertheim JO, Chu DK, Peiris JS, Kosakovsky Pond SL, Poon LL.** 2013. A case
658 for the ancient origin of coronaviruses. *J Virol* **87**:7039-7045.
- 659 49. **Wertheim JO, Pond SLK.** 2011. Purifying Selection Can Obscure the Ancient Age
660 of Viral Lineages. *Molecular Biology and Evolution* **28**:3355-3365.
- 661 50. **Wertheim JO, Smith MD, Smith DM, Scheffler K, Kosakovsky Pond SL.** 2014.
662 Evolutionary origins of human herpes simplex viruses 1 and 2. *Mol Biol Evol*
663 **31**:2356-2364.
- 664 51. **Feltkamp MC, Kazem S, van der Meijden E, Lauber C, Gorbalenya AE.** 2013.
665 From Stockholm to Malawi: recent developments in studying human polyomaviruses.
666 *J Gen Virol* **94**:482-496.
- 667 52. **Streicker DG, Turmelle AS, Vonhof MJ, Kuzmin IV, McCracken GF, Rupprecht**
668 **CE.** 2010. Host phylogeny constrains cross-species emergence and establishment of
669 rabies virus in bats. *Science* **329**:676-679.
- 670 53. **Deuzing I, Fagrouch Z, Groenewoud MJ, Niphuis H, Kondova I, Bogers W,**
671 **Verschoor EJ.** 2010. Detection and characterization of two chimpanzee polyomavirus
672 genotypes from different subspecies. *Virol J* **7**:347.

- 673 54. **Johne R, Enderlein D, Nieper H, Muller H.** 2005. Novel polyomavirus detected in
674 the feces of a chimpanzee by nested broad-spectrum PCR. *J Virol* **79**:3883-3887.
- 675 55. **Madinda NF, Robbins MM, Boesch C, Leendertz FH, Ehlers B, Calvignac-**
676 **Spencer S.** 2015. Genome Sequence of a Central Chimpanzee-Associated
677 Polyomavirus Related to BK and JC Polyomaviruses, Pan troglodytes troglodytes
678 Polyomavirus 1. *Genome Announc* **3**.
- 679 56. **van Persie J, Buitendijk H, Fagrouch Z, Bogers W, Haaksma T, Kondova I,**
680 **Verschoor EJ.** 2016. Complete Genome Sequence of a Novel Chimpanzee
681 Polyomavirus from a Western Common Chimpanzee. *Genome Announc* **4**.
- 682 57. **Peretti A, FitzGerald PC, Bliskovsky V, Pastrana DV, Buck CB.** 2015. Genome
683 Sequence of a Fish-Associated Polyomavirus, Black Sea Bass (*Centropristis striata*)
684 Polyomavirus 1. *Genome Announc* **3**.
- 685 58. **Anthony SJ, Islam A, Johnson C, Navarrete-Macias I, Liang E, Jain K, Hitchens**
686 **PL, Che X, Soloyvov A, Hicks AL, Ojeda-Flores R, Zambrana-Torrel C, Ulrich**
687 **W, Rostal MK, Petrosov A, Garcia J, Haider N, Wolfe N, Goldstein T, Morse SS,**
688 **Rahman M, Epstein JH, Mazet JK, Daszak P, Lipkin WI.** 2015. Non-random
689 patterns in viral diversity. *Nat Commun* **6**:8147.

690

691

692 **Figure legends**

693 Figure 1. Maximum likelihood tree derived from an alignment of partial VP1 sequences. This
694 tree was rooted at its center. The six grey circles stand for the main nodes whose date
695 estimates are given in full in **Tables 5** and **6**; the black circle indicates the node that was used
696 to calibrate the analyses. Note that these circles coincide with putative co-divergence events.
697 This tree was rooted using TempEst. Bp: bootstrap, pp: posterior probability.

698

699 Figure 2. Chronogram derived from an alignment of partial VP1 sequences. This chronogram
700 was obtained through BMCMC analyses run under a multi-species coalescent model (the
701 clades corresponding to entities considered as species are highlighted in blue). Other
702 BMCMC analyses run under different tree priors and ML analyses gave similar results. The
703 root of the tree was the most frequently observed in all posterior samples of trees (pp ca. 0.60)
704 and was also retrieved by rooting the ML tree at its center. The six grey circles stand for the
705 main nodes whose date estimates are given in full in **Tables 5** and **6**; the black circle indicates
706 the node that was used to calibrate the analyses. Note that these circles coincide with putative
707 co-divergence events. Bp: bootstrap, pp: posterior probability.

708

709 **Tables**

710 Table 1. Samples and screening results.

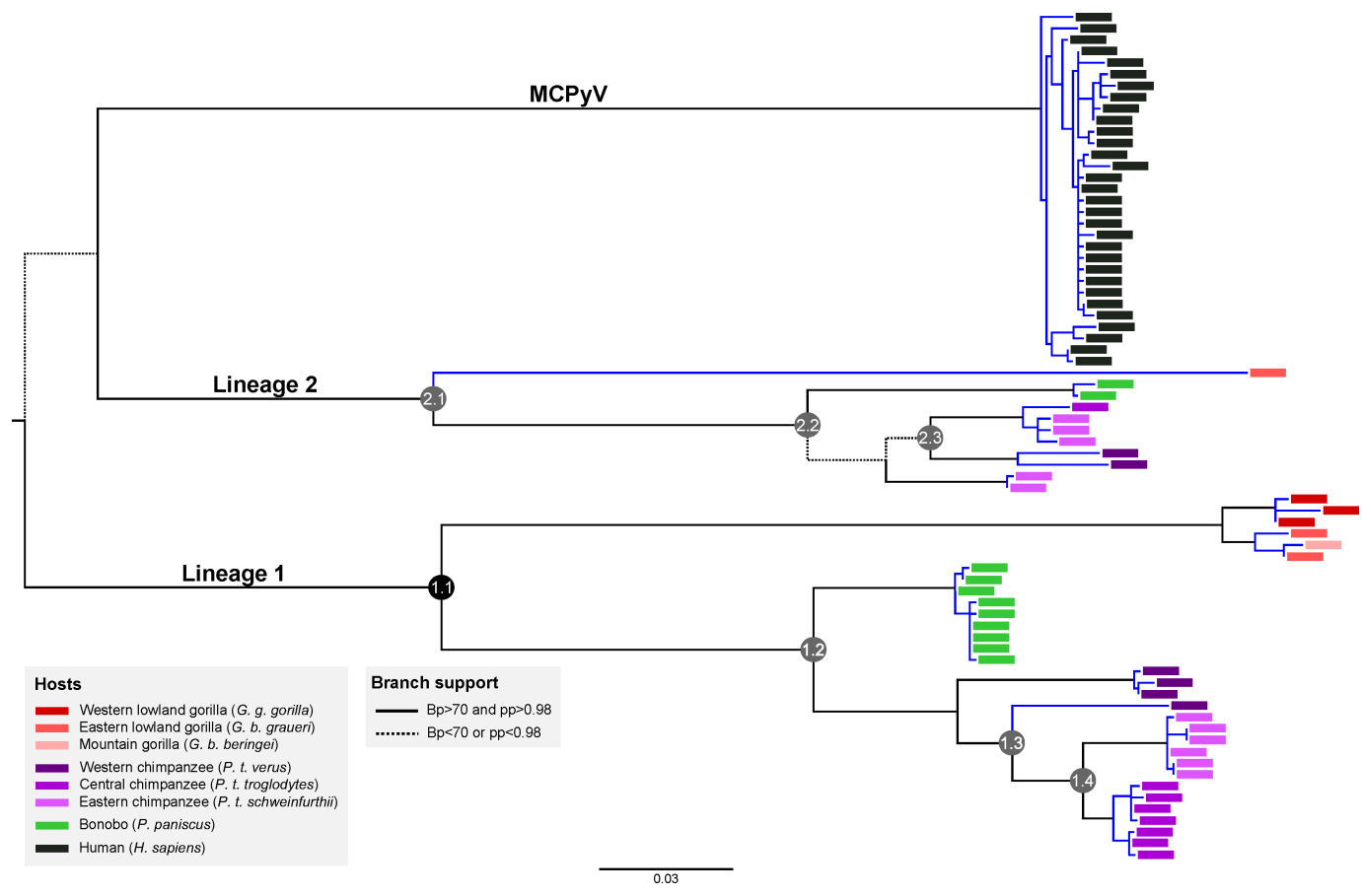
711 Table 2. Primers used in this study.

712 Table 3. Results of BaTS tests for host-specificity.

713 Table 4. Results of reconciliation analyses with Jane.

714 Table 5. Absolute estimates of times to the most recent common ancestors (tMRCA) of PyV
715 in lineages 1 and 2.

716 Table 6. Relative estimates of times to the most recent common ancestors (tMRCA) of PyV in
717 lineages 1 and 2.



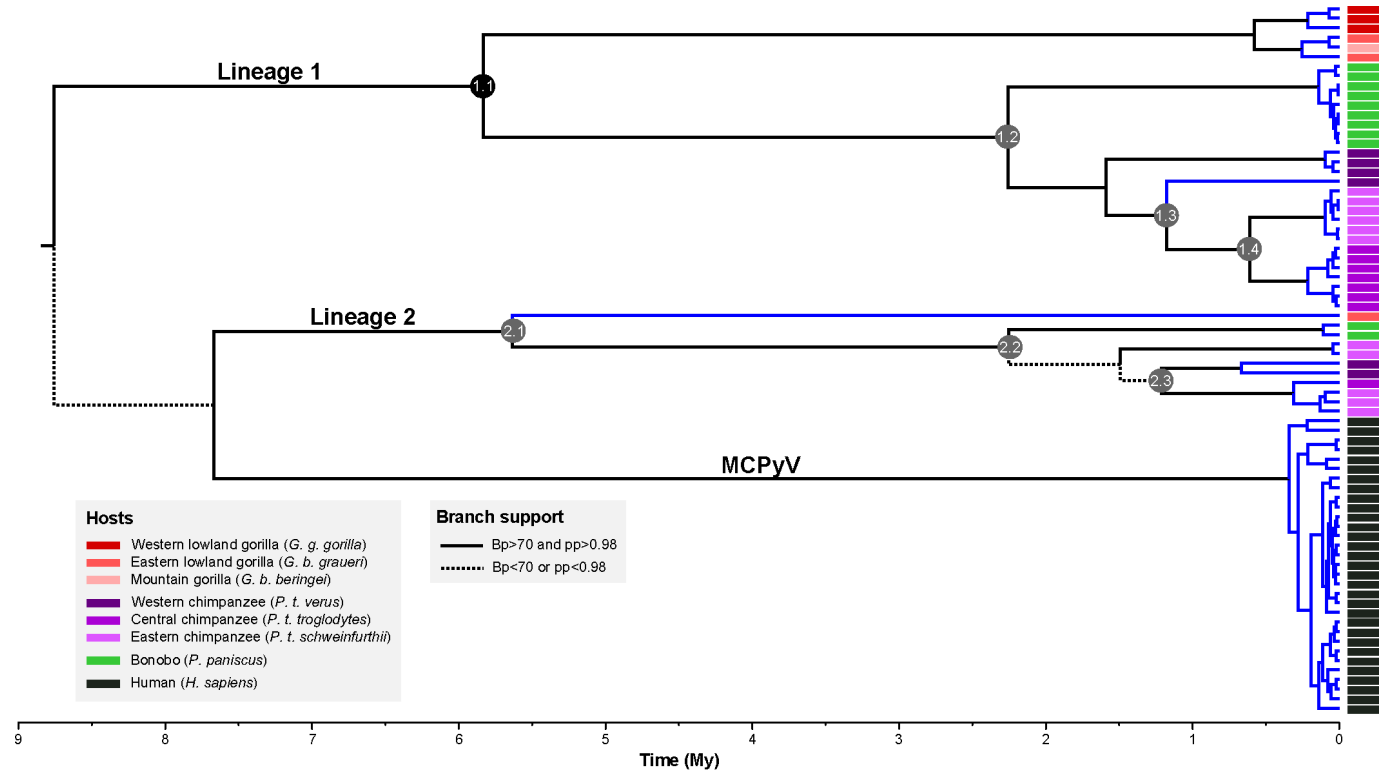


Table 1. Samples and screening results. PyV sequences from Western chimpanzees (*P. t. verus*) were already available from a previous study (14).

Species/sub-species	Country	Site	Samples	Positives	Proportion [95% confidence interval] (%) [*]	Minimum identity within host sub-species (%) [*]	Maximum identity with a publicly available sequence (% accession number, host sub-species)
<i>Gorilla gorilla gorilla</i>	Cameroon	Belgique	20	1			
		Campo Ma'an National Park	18	0			
		Mambele	19	0			
	Central African Republic	Dzanga-Sangha Special Reserve	23	0			
		Loango National Park	25	1	1.9 [0.3-7.4]	98.5	99/HQ385752/ <i>G. g. gorilla</i>
<i>Gorilla gorilla diehli</i>	Cameroon	Greater Takamanda - Mone Landscape	22	0	0 [0-18.5]	NA	NA
<i>Gorilla beringei beringei</i>	Rwanda	Volcanoes National Park	51	0			
	Uganda	Bwindi Impenetrable National Park	30	1	1.2 [0-7.6]	NA	98/HQ385752/ <i>G. g. gorilla</i>
<i>Gorilla beringei graueri</i>	Democratic Republic of the Congo	Kahuzi-Biega National Park	34	7	20.6 [9.3-38.4]	74.7	99/HQ385752/ <i>G. g. gorilla</i> 83/HQ385747/ <i>P. t. verus</i>
<i>Pan troglodytes troglodytes</i>	Cameroon	Belgique	5	1			
		Cameroun	10	1			
		Campo Ma'an National Park	1	0			
		Mambele	9	1			
	Gabon	Loango National Park	27	3	11.5 [4.8-24.1]	77	95/HQ385748/ <i>P. t. verus</i> 94/HQ385747/ <i>P. t. verus</i>
<i>Pan troglodytes schweinfurthii</i>	Uganda	Budongo Central Forest Reserve	33	9			
		Kibale Forest National Park	33	11	30.3 [20-43]	76.5	95/HQ385747/ <i>P. t. verus</i> 94/HQ385748/ <i>P. t. verus</i>
<i>Pan paniscus</i>	Democratic Republic of the Congo	Salonga National Park	26	14	53.8 [33.7-72.9]	77.4	91/HQ385751/ <i>P. t. verus</i> 91/HQ385746/ <i>P. t. verus</i>
<i>Homo sapiens</i>	Côte d'Ivoire	Tai National Park	96	16			
	Democratic Republic of the Congo	Salonga National Park	101	14	15.2 [10.7-21.2]	99	100/JF812999/ <i>H. sapiens</i>

^{*} At the species/sub-species level. NA: not assessed.

Table 2. Primers used in this study.

	Primer name	Primer sequence (5'-3')	Annealing temperature (°C)	Fragment size
PCR1	PCR1.1-f	TGTGCTCCTAAGCCBGGATG	57	
	PCR1.1-r	ACTACTGGGTATGGRTTYTTMACC		
	PCR1.2-f	CTGAATCCAAGRATGGGAGT	58	0.7 kb
	PCR1.2-r	CATGAAANGCCATTTTNCCT		
PCR2	PCR2.1-f	CTGAAGYCTGGGACGMTGAG	57	
	PCR2.1-r	GCAAACATRTGRTAATTGACTCCC		
	PCR2.2-f	TCAGACWCCSAGTCCAGAGG	58	2.5 kb
	PCR2.2-r	GCAAATCYARRGGYTCTCCTC		
PCR3	PCR3.1-f	TGATATGCAGCCMAATMWWCARC	58	
	PCR3.1-r	AAACATGTGATAATTGACTCCCTC		
	PCR3.1-f	TGATATGCAGCCMAATMWWCARC	58	0.2 kb
	PCR3.2-r	AATTGACTCCCTCAATAGGAATG		

Table 3. Results of BaTS tests for host-specificity. The values reported are derived from analyses performed on posterior sets of trees generated under the 14-species coalescent model. Values were very similar when analyzing posterior samples of trees obtained under a constant population size coalescent model or a birth-death speciation model.

Host grouping (# categories)	Mean association index	Mean parsimony score	Mean maximum exclusive single-state clade size	p-value
Species (5)	0.016	6	-	0
<i>Gorilla beringei</i>	-	-	3	<0.002
<i>Gorilla gorilla</i>	-	-	3	<0.002
<i>Homo sapiens</i>	-	-	31	<0.002
<i>Pan paniscus</i>	-	-	9	<0.002
<i>Pan troglodytes</i>	-	-	17	<0.002
Sub-species (8)	0.4	11	-	0
<i>G. b. beringei</i>	-	-	1	1
<i>G. b. graueri</i>	-	-	1	1
<i>G. g. gorilla</i>	-	-	3	<0.002
<i>H. sapiens</i>	-	-	31	<0.002
<i>P. paniscus</i>	-	-	9	<0.002
<i>P. t. schweinfurthii</i>	-	-	6	<0.002
<i>P. t. troglodytes</i>	-	-	7	<0.002
<i>P. t. verus</i>	-	-	3	<0.002

Table 4. Results of reconciliation analyses with Jane.

Host phylogeny	Cost set	Number of events*		p-value
		Co-speciation	Not co-speciation	
Species level	1	5	2	0.056
	2	5	2	0.016
	3	5	2	0.066
Sub-species level	1	10	5	0
	2	10	5	0
	3	10	5	0

* For the solution which was the most parsimonious in number of events.

1 **Table 5. Absolute times to the most recent common ancestors (tMRCA) of PyV in lineages 1 and 2.** Estimates that are incompatible with those
2 determined in Prado-Martinez et al. (2013) appear in bold.
3

		Time to the most recent common ancestor (in million years)						
		Median or ML estimate [95% HPD or Bp interval [§]]						
Statistical framework	Diversification model or smoothing factor*	Lineage 1				Lineage 2		
		Node 1.1 [§] (all)	Node 1.2 (panine)	Node 1.3 (<i>P. troglodytes</i>)	Node 1.4 (<i>P.t.t.+P.t.s.</i>)	Node 2.1 (all)	Node 2.2 (panine)	Node 2.3 (<i>P. troglodytes</i>)
BMCMC	Coalescent:		2.15	1.09	0.57	5.36	2.12	1.11
	Constant population size	5.62	[1.54-2.85]	[0.74-1.48]	[0.36-0.85]	[3.71-7.31]	[1.47-2.92]	[0.75-1.55]
	Multi-species coalescent	5.62	2.25	1.18	0.61	5.63	2.25	1.21
	Speciation:		[1.57-3.11]	[0.79-1.67]	[0.37-0.91]	[3.84-7.95]	[1.52-3.18]	[0.79-1.73]
ML	Birth-death	5.62	2.06	1.05	0.54	5.27	2.04	1.07
			[1.46-2.72]	[0.71-1.44]	[0.34-0.79]	[3.58-7.20]	[1.38-2.80]	[0.69-1.48]
	1	5.62	2.29	1.06	0.54	5.51	2.27	1.21
			[1.61-4.56]	[0.76-2.22]	[0.23-0.84]	[3.97-24.92]	[1.77-25.37]	[0.75-13.33]
Prado-Martinez et al. (2013) [¶]	100	5.62	2.24	1.04	0.53	5.54	2.26	1.21
			[1.61-4.19]	[0.70-2.27]	[0.18-1.93]	[4.02-9.84]	[1.79-4.98]	[0.76-2.24]
		5.62	0.87	0.42	0.17	5.6	0.87	0.42

4 * Diversification models for BMCMC, smoothing factors for ML (under the MG94-like model of codon evolution).
5 [§] The according node was used to calibrate the trees.
6 [§] 95% HPD for BMCMC, Bp intervals for ML. Bp intervals were determined using 100 bootstrap pseudo-replicates of the codon dataset from which branch lengths were re-
7 estimated on the ML topology; all trees were rooted using TempEst.
8 BMCMC: Bayesian Markov chain Monte Carlo, ML: maximum likelihood, HPD: highest posterior density, Bp: bootstrap.
9 [¶] Assuming a mutation rate of 1e-9 mutation/(bp.y).

1 **Table 6. Relative times to the most recent common ancestors (tMRCA) of PyV in lineages 1 and 2.** Estimates that are incompatible with those
2 determined in Prado-Martinez et al. (2013) appear in bold.
3

		Time to the most recent common ancestor (1 unit=tMRCA of <i>P.t.t.t.</i> + <i>P.t.s.</i>)						
		Median or ML estimate [95% HPD or Bp interval [§]]						
Statistical framework	Diversification model or smoothing factor*	Lineage 1				Lineage 2		
		Node 1.1 [§] (all)	Node 1.2 (panine)	Node 1.3 (<i>P. troglodytes</i>)	Node 1.4 (<i>P.t.t.t.</i> + <i>P.t.s.</i>)	Node 2.1 (all)	Node 2.2 (panine)	Node 2.3 (<i>P. troglodytes</i>)
BMCMC	Coalescent:		3.77	1.91	1	9.40	3.72	1.95
	Constant population size	9.82	[2.70-5.00]	[1.30-2.60]	[0.63-1.49]	[6.51-12.82]	[2.58-5.12]	[1.31-2.72]
	Multi-species coalescent	9.18	3.69	1.93	1	9.23	3.69	1.98
	Speciation:		[2.57-5.10]	[1.29-2.74]	[0.61-1.49]	[6.29-13.03]	[2.49-5.21]	[1.29-2.84]
ML	Birth-death	10.37	3.81	1.94	1	9.76	3.78	1.98
			[2.70-5.04]	[1.31-2.67]	[0.63-1.46]	[6.63-13.33]	[2.56-5.18]	[1.28-2.74]
	1	10.37	4.24	1.96	1	10.20	4.20	2.24
			[2.98-8.44]	[1.41-4.11]	[0.43-1.58]	[7.35-46.15]	[3.28-46.98]	[1.39-24.68]
Prado-Martinez et al. (2013)	100	10.60	4.22	1.96	1	10.45	4.26	2.28
			[3.04-7.90]	[1.32-4.28]	[0.34-3.64]	[7.58-18.57]	[3.38-9.40]	[1.43-4.22]
		32.11	4.98	2.39	1	32.11	4.98	2.39

4 * Diversification models for BMCMC, smoothing factors for ML (under the MG94-like model of codon evolution).
5 [§] No HPD or Bp interval because this node was used to calibrate the trees.
6 [§] 95% HPD for BMCMC, Bp intervals for ML. Bp intervals were determined using 100 bootstrap pseudo-replicates of the codon dataset from which branch lengths were re-
7 estimated on the ML topology; all trees were rooted using TempEst.
8 BMCMC: Bayesian Markov chain Monte Carlo, ML: maximum likelihood, HPD: highest posterior density, Bp: bootstrap.